PB 169 301

RADIONUCLIDE ANALYSIS OF GAMMA-RAY SPECTRA BY STEPWISE MULTIPLE REGRESSION

Robert H. Shumway

Public Health Service
Washington, D. C.

1963

CLEARINGHOUSE
FOR FEDERAL SCIENTIFIC AND TECHNICAL INFORMATION

Distributed . . . 'to foster, serve and promote the nation's economic development and technological advancement.'

U.S. DEPARTMENT OF COMMERCE/National Bureau of Standards

This document has been approved for public release and sale.

PB189361

ENVIRONMENTAL HEALTH SERIES

Radiological Health

RADIONUCLIDE ANALYSIS OF GAMMA-RAY SPECTRA BY STEPWISE MULTIPLE REGRESSION

U. S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

Public Health Service

Reproduced by the
CLEARINGHOUSE
for Federal Scientific & Technical
Information Springfield Va. 22151

ENVIRONMENTAL HEALTH SERIES

This series of reports embraces the science and technology of creating and maintaining a healthful environment for man. Individual reports in the series deal with health factors in: air; water; food; chemical, biological, and radiological contamination; sound, vibration, and x-radiation; liquid, solid, and gaseous wastes; homes, places of employment and recreation, and institutions.

The principal subject area covered by each report in the series is identified by a characteristic letter appearing after the Public Health Service Publication No. "999," designating the series. The following letters and subject areas apply:

A - Air
G - General Engineering
H - Arctic Health
M - Milk and Food
O - Occupational Health
R - Radiological Health
W - Water

Separate Environmental Health Series reports in each subject area are identified by a serial number following the subject area letter.

RADIONUCLIDE ANALYSIS OF GAMMA-RAY SPECTRA BY STEPWISE MULTIPLE REGRESSION

by

Robert H. Shumway

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Public Health Service
Division of Radiological Health
Washington, D.C. 20201

Library of Congress catalog card number 63-62203

Public Health Service Publication No. 999-R-5

UNITED STATES GOVERNMENT PRINTING OFFICE, WASHINGTON: 1963

FOREWORD

The rapid development of technology has effected a concurrent increase in highly sophisticated diagnostic instruments and scientific computers in academic, industrial, and governmental laboratories. To achieve maximum production from these devices, it is becoming common practice to correlate their performance.

In radiological health activities, the linking of radiation-measuring instruments with digital computers can increase the production of data. This report presents an approach to the problem of processing the voluminous data from a multichannel gamma spectrometer through a high-capacity, scientific computer to obtain concentrations of individual radionuclides in an environmental sample. Other approaches are also being studied by this Division as well as by other investigators. It is hoped that this report will be useful to those working on the problem of analyzing gamma-emitting solutions.

Mr. Shumway's application to this subject began during his service as a member of the Division's Biometrics Unit under the direction of Edward S. Weiss, Chief.

Donald R. Chadwick, M. D.
Chief, Division of Radiological Health

ABSTRACT

A modification of the usual multiple regression method is described for use in analyzing a gross gamma spectrum from an unknown solution containing many radionuclides. It involves a stepwise multiple regression procedure that sequentially tests the effect of each radionuclide until the spectrum is resolved. The principal advantages over the "spectrum stripping" or conventional multiple regression are that the procedure: (a) requires no a priori assumptions about the radionuclide content of the solution, and (b) utilizes the information in all of the channels simultaneously.

RADIONUCLIDE ANALYSIS OF GAMMA-RAY SPECTRA BY STEPWISE MULTIPLE REGRESSION

Introduction

Continuing programs involving the assessment of gamma-emitting radionuclides in the environment have created a need for analytical procedures that quickly and accurately measure the radionuclide composition of different samples of food and water. Multichannel gamma-ray spectrometers are used to rapidly analyze a sample for the common gamma emitters, such as iodine-131, cesium-137, barium-140, potassium-40, and others (see references 1-5). However, the errors introduced by the spectrometer (nonlinearity, channel drift, nonuniformity, interfering photopeaks, and so forth) plus the very low levels encountered in the sample combine to make the estimates subject to considerable variability. In addition, it is frequently not known exactly which radionuclides are present so that some method of visual inspection of the data must be made. Presently, a number of different methods are available for dissecting the composite spectrum of a radioactive sample into its component parts when the identity of the contributors is known. Most of these procedures lead to estimates for the concentrations of the different gamma emitters by one of the three following schemes.

One may eliminate successively the elements with the higher energy photopeaks until the entire composite spectrum has been resolved. The successive subtractions (referred to as "stripping") may be accomplished on either analog or digital equipment (1, 2). In most of the variants of gamma spectrum stripping, only the counts in the region of the

photopeak are used. In some cases, estimates can become distorted when there are interfering photopeaks, or when counts from a lower photopeak radionuclide interfere with the calculation of the higher photopeak radionuclide. However, the main objection to such a procedure lies in the error assumptions. Since errors are additive, successive subtractions introduce much higher variance in the radionuclides with lowest energy photopeaks. Thus it is conceivable that the error in the estimate of the last radionuclide might be much larger than the estimate itself.

A second major approach sets up simultaneous equations for n radionuclide concentrations in terms of n counts taken in the photopeak regions (3). This approach again utilizes only the information in photopeak channels but can estimate concentrations when interfering photopeaks are present. However, one gets no valid estimates for error in the concentrations. In addition, the identity of the contributors must be inferred from the composite spectrum before the method is applied.

A method that is gaining increasing favor in the literature (4) is the estimation of the unknown concentrations by least squares. Here it is possible to estimate concentrations over a wide variety of possible situations. The linearity of the hypothesis and the assumption of normally distributed residuals allow one to compute confidence intervals for the concentrations. Furthermore, the estimates have the merit of being efficient and unbiased. However, given an unknown solution of a number of radionuclides, one still must guess a priori at the exact composition of the mixture. For example, one is frequently given a composite spectrum which may contain any, one, or all of a number of radionuclides. The application of conventional weighted or unweighted least squares requires that a subset of assumed contributors be selected and the corresponding analysis performed. Obviously, it is not feasible to do multiple regressions for every possible subset of radionuclides. Thus the investigator is frequently reduced to visual analysis of the spectrum to guess at appropriate nuclides which might be present.

The preceding discussion emphasizes the need for a more sophisticated program which can take a spectrometer reading from an unknown combination of low-level gamma-emitting radionuclides and rapidly determine the concentration of each one present with confidence limits closely bracketing the estimated concentration. The approach considered here will

be in terms of a sequential weighted multiple regression analysis. It represents an attempt to integrate a decision-making process into the standard weighted least squares analysis. In this respect it could be considered as a combination of the stripping and least squares approaches, although the decisions are influenced by the statistical F tests in the analysis of variance rather than by the existence of a photopeak. In addition, the estimated errors at the end of the decision process are those which would prevail in a least squares solution with the stated radionuclides present. The stepwise multiple regression approach is completely general in that it could theoretically handle an unknown solution containing any, one, or all of up to 60 radionuclides, if a library of 60 reference spectra were available. However, limitations in the composition of the spectra would probably preclude the use of the analysis for such a large number of radionuclides. For example, several radionuclides may have identical, or nearly identical, spectra. This is only one of a number of practical limitations which will be discussed in the next two sections.

Theoretical Development

All of the approaches currently in use in gamma spectrometry assume that the gross counts in the composite spectrum are linear combinations of the counts in the reference spectra, the weights being proportional to the radionuclide concentrations. That is, in matrix motion

$$y = Xc + e \qquad (1)$$

where $y' = (y_1, y_2, \ldots, y_n)$ is a vector whose i^{th} element is the gross count of an unknown sample in the i^{th} channel of a spectrometer and c is the unknown vector of concentrations (of k radionuclides) where $c' = (c_1, \ldots, c_k)$. The rectangular array $X = (x_{ij})$ is the reference matrix made up from the reference standards of the pure radionuclides, x_{ij} being the contribution of the j^{th} radionuclide present in the i^{th} channel. The expression $e' = (e_1, \ldots, e_n)$ is a random error vector with $E(ee') = \sum$ as its variance-covariance matrix. The differences in the present approaches stem from the different methods for estimating the concentration vector c and the error in the resulting concentrations. In addition, an ideal approach must provide a means for deciding which radionuclides are present in the unknown solution.

This may be predetermined by inspecting the spectra or may be inferred as the fitting procedure progresses.

Suppose, now, that the gross counts are expressed as linear combinations of the reference spectra in Eq. (1). The least squares solution for the concentrations is:

$$c = (X'\sum^{-1}X)^{-1}X'\sum^{-1}y \qquad (2)$$

Assume that an unknown solution containing between 1 and n nuclides is given. The stepwise analysis used here, as developed by Efroymson (5), is a method for sequentially testing each radionuclide to determine which ones give significant improvement in the "goodness of fit" of the observed to the predicted spectrum.

The least squares equations are completely recalculated each time a new radionuclide is considered and all decisions to enter radionuclides are dependent upon the equation obtained up to that time. Instead of entering a new radionuclide, the decision can be made to eliminate a radionuclide already entered. This can occur at any time after two radionuclides have been entered into the equation. The bases upon which a decision for entering or removing nuclides is made are familiar F statistics. The F levels for entering or removing radionuclides are arbitrarily selected before the analysis is performed. The following discussion, taken from Coons and Falter (6), illustrates the procedure.

Of those radionuclides already entered, the machine identifies that one, x_j, which is associated with the minimum additional sum of squares of regression after removing that explained by the other entered radionuclides. The F ratio $\min j/s^2_{y.12\ldots j\ldots m}$, where $s^2_{y.12\ldots j\ldots m}$ is the unexplained variance of y after having fit variables $1, 2, \ldots j, \ldots m$ in the equation, is compared with the selected level for removing radionuclides. If it is greater, x_j remains in the equation. Otherwise, it is eliminated. If the radionuclide is removed, then the equation and matrix are adjusted accordingly and the whole process of deciding whether to remove or enter a radionuclide is repeated, given the new reduced equation. If the decision is made to let the radionuclide remain in the equation, then the machine selects a new radionuclide. That is, of the radionuclides not yet entered, the machine identifies that one, x_i, which is associated with the maximum additional sum of squares explained by regression, given that

the other k radionuclides are present. Designate this maximum additional sum of squares by max_i and form the ratio $max_i / s^2_{y.12...i...m}$. Compare this ratio with the selected F level to enter radionuclides in the regression. If x_i is entered, then the regression equation and matrix are adjusted accordingly and the cycle is repeated. If x_i is rejected, the process is terminated.

The sequential testing method might seem to involve an inordinately large number of matrix inversions. However, the fact of the matter is that the Gaussian elimination method used yields the partial regression coefficients at each step of the procedure merely by a simple sequence of row and column operations. When a radionuclide is entered, then the row and column corresponding to that radionuclide are used as pivotal elements, and an adjustment procedure similar to the Doolittle technique (7) is carried out. If a radionuclide is removed from the equation, then the row and column corresponding to that radionuclide are again used as pivotal elements in the same operations so that the inverse matrix returns to the state it was in before the nuclide was entered.

Now that the radionuclides have been specified by means of a reasonable decision scheme, some sort of weighting scheme must be used. From Eq. (2) it can be seen that this is equivalent to specifying a form for the variance-covariance matrix $\sum$. If the errors are uncorrelated with zero mean and constant variance we have the usual weighted analysis where the weights are the inverses of the variances. Another simple assumption is that the errors are uncorrelated with zero mean and different variances $k_i \sigma^2$. If the gross counts are Poisson random variables, then $k_i \sigma^2 = y_i$, a procedure which weights the channel according to the count rate (4). Another possibility is to weight more heavily in photopeak regions so that the fitted function agrees more closely at the photopeak values. For example, one might take the weights to be one in the photopeak regions and zero elsewhere, in which case there would obtain n equations in n unknowns with $\sum = I$, and the least squares solution would become

$$c = (X'X)^{-1}X'y = X^{-1}y$$

which is equivalent to the simultaneous equation approach of Hagee, Karches and Goldin (3).

The stepwise procedure as considered above is open to question. For example, what happens when two spectra are

identical or, at least, linear combinations of each other? The above procedure, as well as the others in use, depends upon the linear independence of the component spectra. However, unlike the stripping and simultaneous equations approaches which break down when photopeak counts are proportional, this procedure breaks down only when the counts in all channels are exactly proportional. In this unlikely case the program eliminates the radionuclides that are linear combinations of the maximum contributor, and one must make a practical decision as to which radionuclides are present. Any least squares program would completely break down in this singular case. This situation is very unlikely and does not arise in any of the computed examples. A second major objection is the order in which the tests are performed. As the number of radionuclides is increased it is possible that the program might not lead to the least squares solution. However, the examples computed here give results in terms of F ratios which are significant at considerably higher than the 99.99% level. In addition, it is possible that the least squares solution is not identical with the fit that would be obtained using the real components of the composite spectrum. In this connection it may be argued that a useful way to analyze a solution would be to assume that all nuclides are present and hope that those not present would give zero estimates for the concentrations. When the number of radionuclides is small this method, although logically inconsistent, might produce reasonable estimates. However, as the number of radionuclides increases the chances for spurious linear combinations also are increased, and inaccurate and misleading estimates for the concentrations might be obtained.

In the next section a series of examples is presented that illustrate the efficiency with which the stepwise procedure copes with a number of different problems. In order to illustrate the power of the stepwise method, a number of contrived solutions are used which present some commonly occurring problems. The solutions are contrived in the sense that they represent known linear combinations of the radionuclides plus a random error. The use of contrived solutions is a convenient procedure in that the exact concentrations of the solutions which are being analyzed are known. Thus the results of different estimation procedures can be fairly evaluated. In addition, one can construct solutions which present difficulties to the analysis procedure. For example, in the next section contrived "unknown" solutions will be considered which contain combinations of many nuclides, few nuclides,

and nuclides with interfering, completely obscured, or false "photopeaks". The property of the stepwise multiple regression analysis which enables it to treat these difficult cases so successfully is that it uses the information contained in every channel to estimate concentrations, at each step of the procedure.

Examples

In order to test the stepwise procedure in different situations, the library of reference spectra shown in table 1 in the appendix was used to make up a number of different solutions. The "contrived" solutions shown in table 1 were constructed by taking linear combinations of the reference spectra and adding a random error with a variance of 100 counts in each channel. The composite spectra for the solutions are shown in the appendix (figures 1-5). The basic inputs to the program are taken to be the composite spectra and the library of reference spectra. Any, one, or all of the radionuclides may be present and it is required to specify for each "unknown" spectrum the component concentrations and their standard errors. The first example illustrates the procedure for a commonly occurring simple unknown spectrum.

Commonly Occurring Combination

In solution A, figure 1, the four radionuclides present are easily identified by their distinctive photopeaks. In the second frame (B) of figure 1, the effect of adding the most significant contributor can be seen. The least squares fit is in terms of zinc-65 only. The next most significant F ratio leads to the second step results shown in frame C where cesium-137 has entered at an F value of 1790 which represents a significance of greater than 99.99%. At this point the deletion test fails to eliminate zinc-65. The program continues on to the five-step solution shown in frame F where all nuclides and background have been added at greater than the 99.99% significance. The final estimates for the concentrations are shown in table 2.

Simple Combination

In solution B, figure 2, the ability of the program to correctly specify a small subset of the original library is tested. Note the "false photopeak" at 110 which could be misinterpreted as zinc-65 when actually it is a secondary peak of cobalt-60. Note also the proximity of the peaks, although this problem will be tested more severely in other components given in table 1.

Combination with Interfering Photopeaks

Solution C, figure 3, contains four radionuclides, two of which are ruthenium-rhodium-106 and barium-lanthanum-140 containing peaks at 50 and 48, respectively. Table 2 shows that the precision of the results is still not adversely affected.

Complex Combination with Completely Obscured Photopeak and False Photopeak

In solution D, figure 4, the iodine-131 photopeak at channel 36 is completely obscured because of the small amount of iodine-131 present (.5 spectra units). However, the interesting anomaly is the completely false photopeak found at channel 51 implying (as in the previous solution) the presence of either ruthenium-rhodium-106 or barium-lanthanum-140 or both. The stepwise program correctly identifies this peak as a secondary contribution of zinc-65 and the precision of the results is excellent.

Complex Combination with Completely Obscured Photopeak

In solution E, figure 5, a complex mixture containing seven radionuclides and background has been analyzed. Again the solution contains iodine-131 obscured. Even in this complex case the contributors assert themselves at probability levels considerably in excess of 99.99% and have confidence intervals for component radionuclide concentrations closely bracketing the known values. Summaries of the results of solutions A-E in table 2 demonstrate the potential power of the stepwise procedure for dissecting complex solutions. The application of this method to actual non-contrived solutions would present only the difficulty of estimating the structure of the variance-covariance matrix so that the optimal weighting factors could be applied. Also, there is no difficulty, in principle, to extending the methods to solutions which may contain up to 60 radionuclides, although the procedure for adding and deleting contributors may occasionally break down.

Summary

It has been shown that a modification of the usual multiple regression procedure can be used to analyze a gross gamma spectrometer reading from an unknown solution containing many radionuclides. The modification is based upon a stepwise multiple regression method which sequentially tests the effect of each radionuclide until the contents of the solution are completely specified. The principal advantages of such a method over the usual "spectrum stripping" or conventional multiple regression are that it: (a) requires no a priori assumptions about the radionuclide content of the solution, and (b) utilizes the information in all of the channels simultaneously. Statistical estimates for the errors in the resulting concentrations are provided.

ACKNOWLEDGMENTS

The author wishes to thank his present employers, United ElectroDynamics, Inc., for their generosity in making available computer facilities and related services which made possible the completion of this report.

REFERENCES

1. Folsom, T. R. and R. A. Cramer: "The Use of Magnetic-Tape Loops for Storage and Evaluation of Gamma-Ray Spectra," in Whole-Body Counting, Proceedings Series, International Atomic Energy Agency, Vienna (1962), pp. 367-374.

2. Gustafson, P. F., L. D. Marinelli, and S. S. Brar: "Natural and Fission-Produced Gamma-Ray Emitting Radioactivity in Soil," Science 127:1240-1242 (1958).

3. Hagee, G. R., G. J. Karches, and A. S. Goldin: "Determination of ^{131}I, ^{137}Cs, and ^{140}Ba In Fluid Milk by Gamma Spectroscopy," Talanta 5:36-43 (1960).

4. Pasternack, B. S.: "Linear Estimation and the Analysis of Gamma Ray Pulse Height Spectra," Technometrics 4:565-572 (1962).

5. Efroymson, M. A.: "Multiple Regression Analysis," in Mathematical Methods for Digital Computers (A. Ralston and H. S. Wilf, Eds.), part V, (17), John Wiley & Sons, Inc., New York City (1960), pp. 191-203.

6. Coons, I. J., and K. H. Falter: "Discussion and Verification of Stepwise Multiple Linear Regression Program," C-A-S-E Report 59-2011 Westinghouse Electric Corp. (August 5, 1959).

7. Snedecor, G. W.: "Statistical Methods," Iowa State College Press, Ames, Iowa (1946).

APPENDIX

Table 1.--Reference Library and "Contrived" Solutions

Reference Spectra	Photopeak	Solution A	Solution B	Solution C	Solution D	Solution E
Background		1.00	1.00	1.00	1.00	1.00
$Ru\text{-}Rh^{106}$	50			2.00		2.00
$Zr\text{-}Nb^{95}$	75					3.00
Zn^{65}	109	.25		4.00	2.00	3.50
Co^{60}	130		1.00	1.00		3.00
Cs^{137}	65	1.00			1.00	1.50
$Ba\text{-}La^{140}$	48			1.50		
K^{40}	142	.75	1.10			2.00
I^{131}	36	1.50			.50	1.00

Solution A - Commonly occurring combination

Solution B - Simple combination

Solution C - Combination with interfering photopeaks

Solution D - Combination with a completely obscured photopeak and a false photopeak

Solution E - Complex combination with completely obscured photopeak and interfering photopeaks

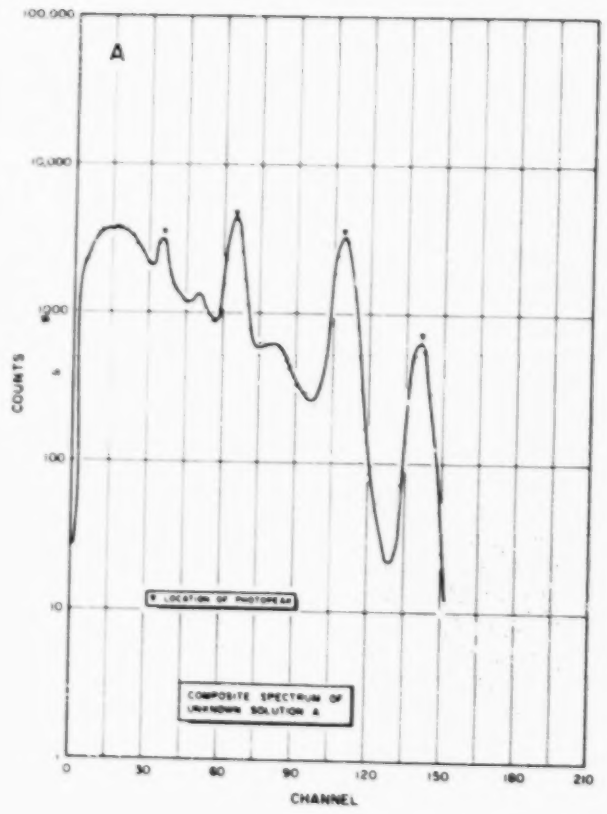
Table 2.--Estimated Concentrations and Associated F Values

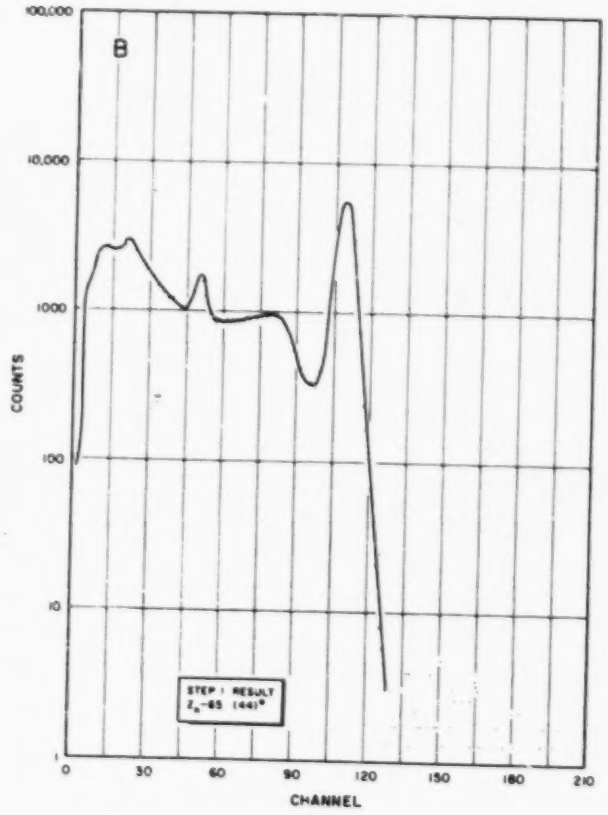
Reference Spectra	Solution A	Solution B	Solution C	Solution D	Solution E
Background	.929-1.013* F=2149**	.938-1.008 F=3023	.913-1.013 F=1490	.929-1.013 F=2166	.998-1.002 F=817
$Ru\text{-}Rh^{106}$			1.998-2.086 F=1676		1.973-2.045 F=12015
$Zr\text{-}Nb^{95}$					2.997-3.001 F=827
Zn^{65}	.2495-.2506 F=639		3.999-4.001 F=11385	1.999-2.001 F=19087	3.499-3.501 F=3272
Co^{60}		.999-1.009 F=1486	.999-1.013 F=1155		2.997-3.009 F=169
Cs^{137}	.998-1.002 F=1791			.998-1.002 F=11967	1.497-1.501 F=244
$Ba\text{-}La^{140}$			1.497-1.501 F=15952		
K^{40}	.741-.760 F=2149	1.091-1.109 F=3365			1.992-2.008 F=367
I^{131}	1.498-1.520 F=739			.498-.520 F=1329	.996-1.012 F=1242

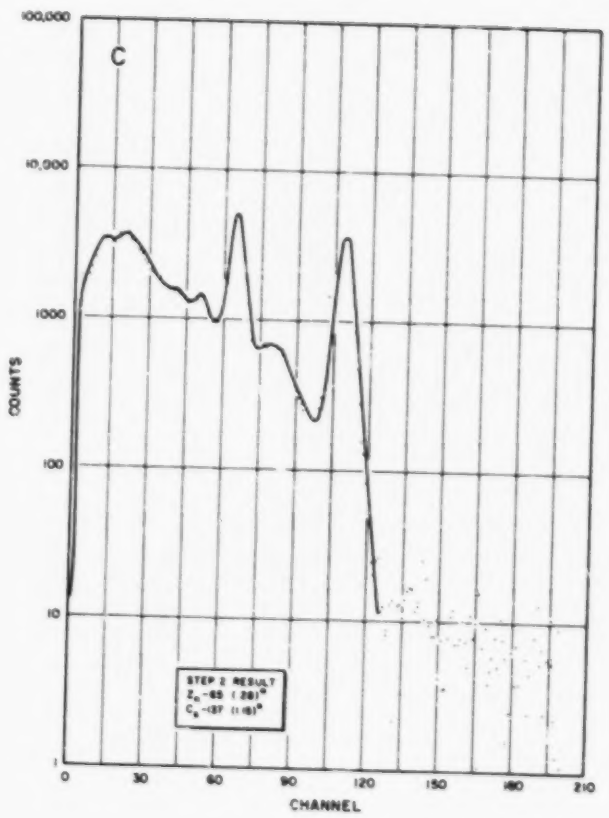
*95% confidence limits in reference spectra units

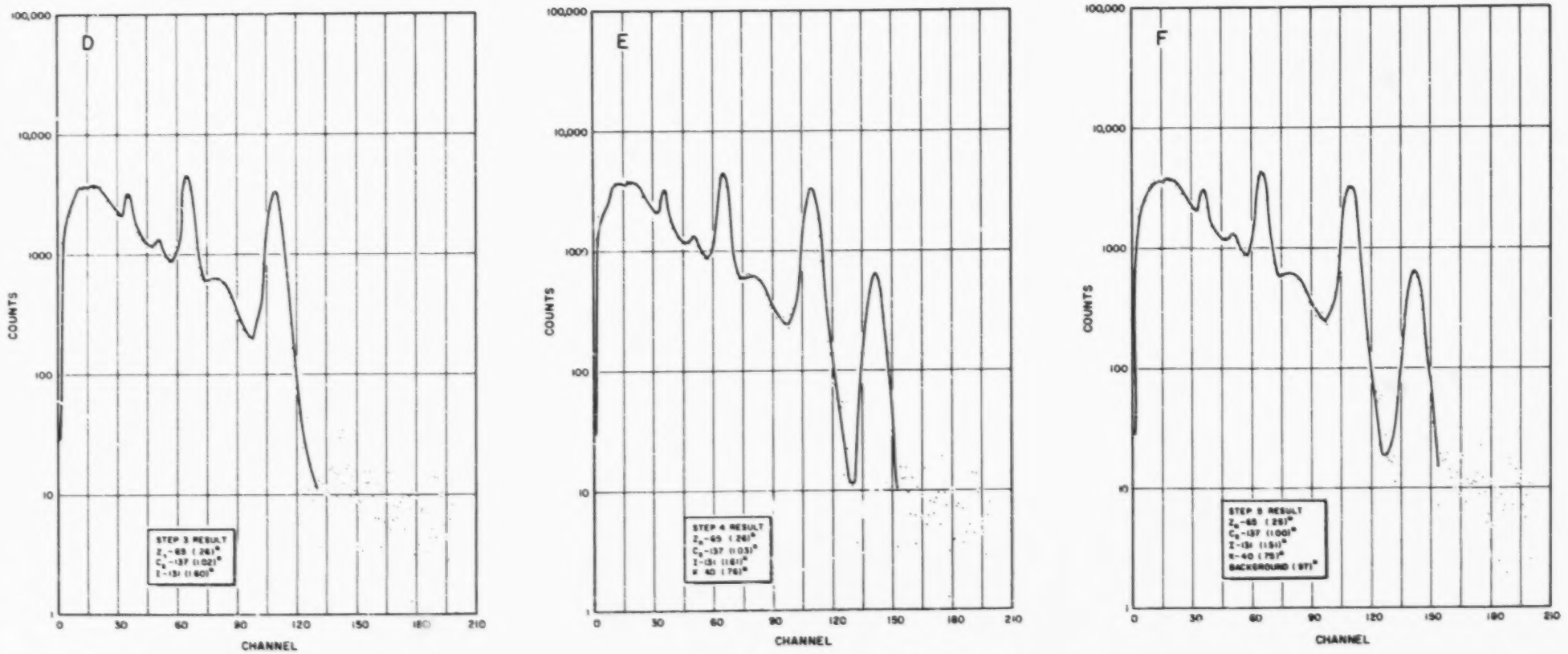
**Minimum F value for entering a nuclide is 4.5; Maximum F value for deleting a nuclide is .5

Illustration of Stepwise Analysis of a Commonly Occurring Solution of Radionuclides









*( ) ESTIMATED CONCENTRATION

Figure 1.--Solution A

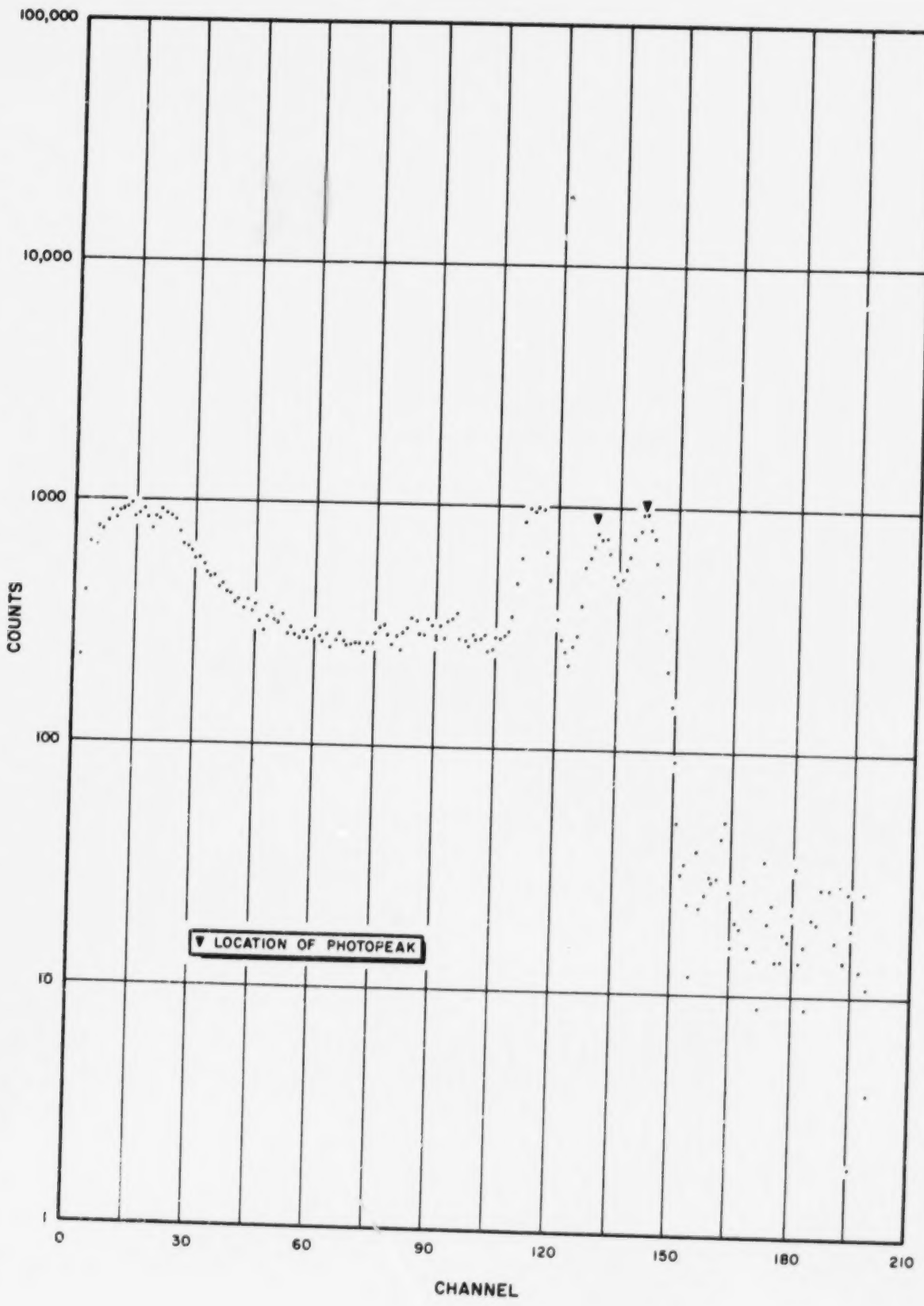


Figure 2.--Unknown Solution B

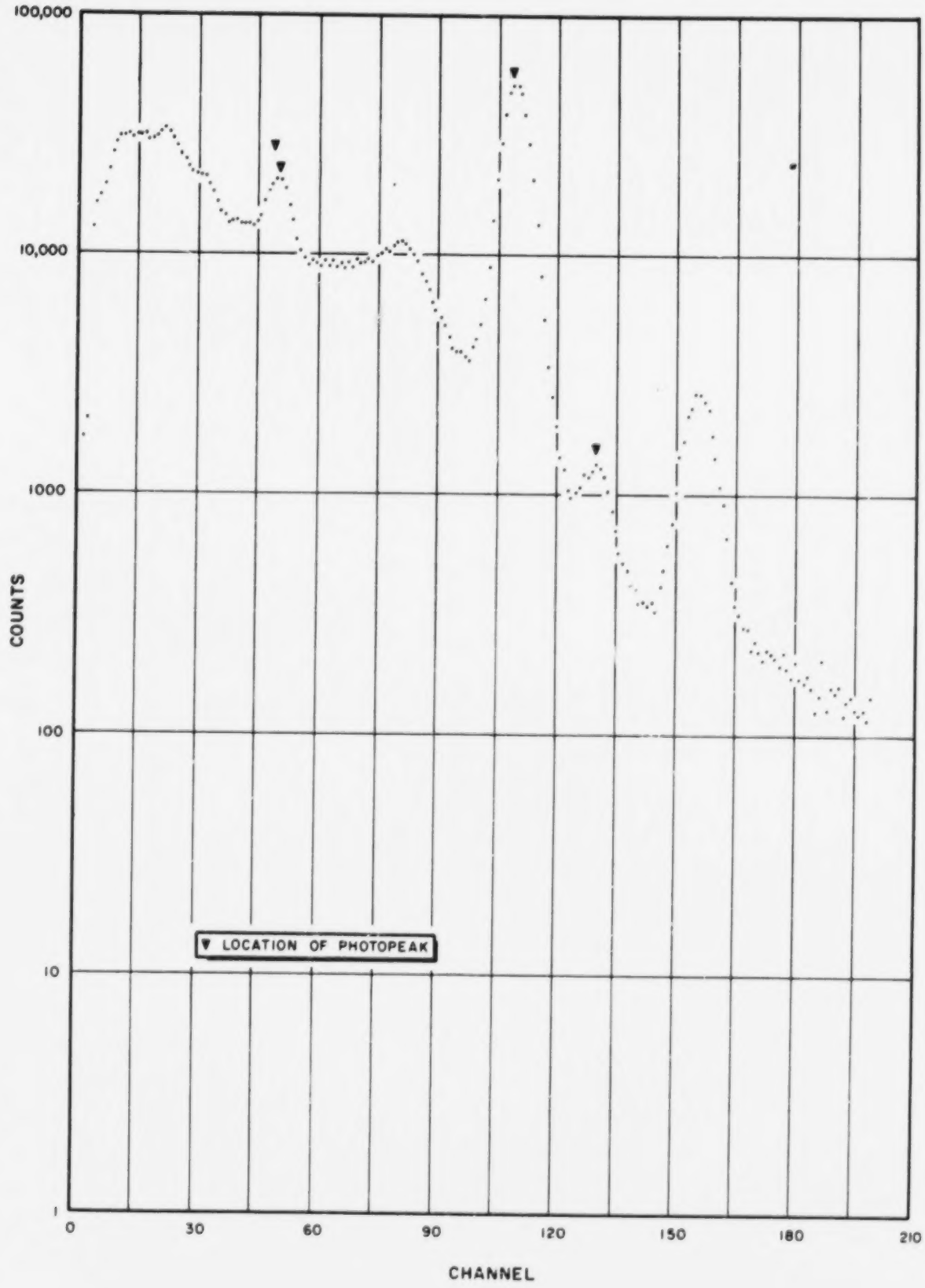


Figure 3.--Unknown Solution C

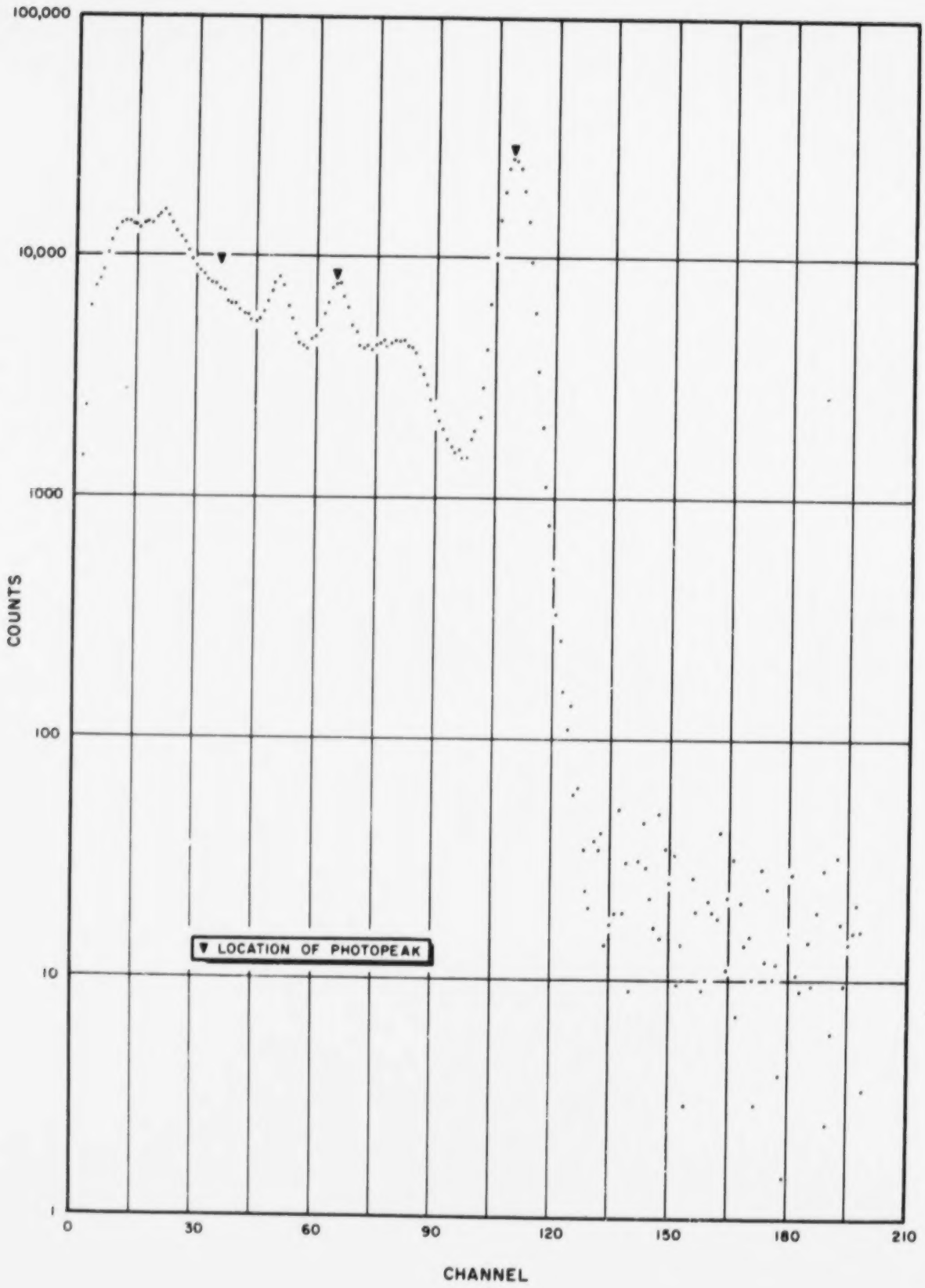


Figure 4.--Unknown Solution D

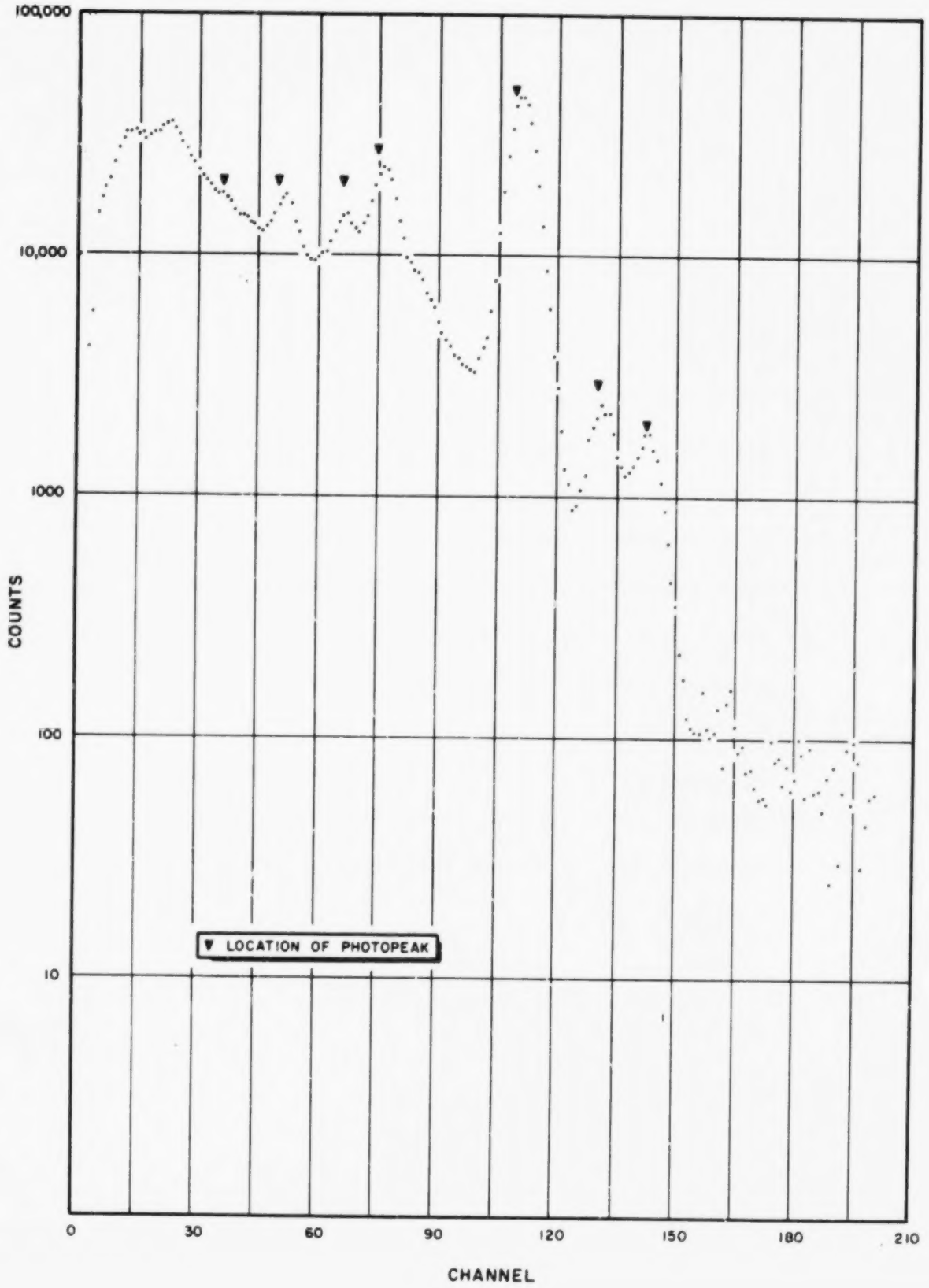


Figure 5.--Unknown Solution E

THE ABSTRACT CARDS accompanying this report are designed to facilitate information retrieval. They provide space for an accession number (to be filled in by the user), suggested key words, bibliographic information, and an abstract. The key word concept of reference material filing is readily adaptable to a variety of filing systems ranging from manual-visual to electronic data processing. The cards are furnished in triplicate to allow for flexibility in their use.

* U.S. GOVERNMENT PRINTING OFFICE 1964 O—715-653